Module 11: Machine Learning

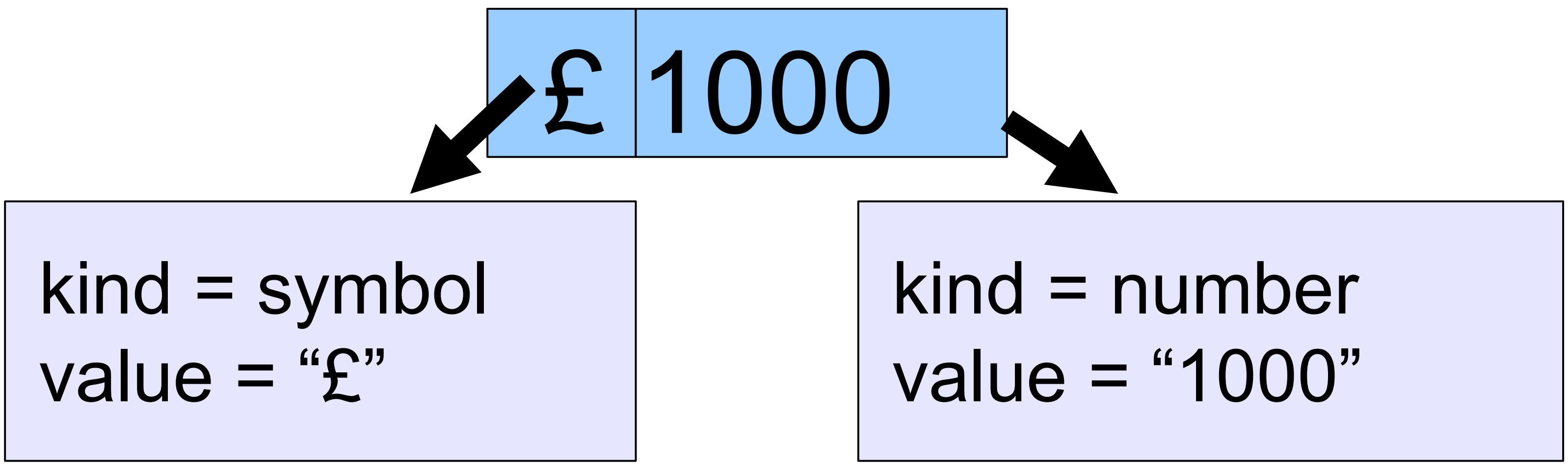
What is Machine Learning and why do we want to do it?

What is ML?

- Automating the process of inferring new data from existing data
- In GATE, that means creating annotations by learning how they relate to other annotations

Learning a pattern

- For example, we have "Token" annotations with "kind" and "value" features



- ML could learn that a "£" followed by a number is an amount of currency

How is that better than making rules (e.g. JAPE)?

- It is different to the rule-based approach
- Some things humans are better at writing rules for, and some things ML algorithms are better at finding
- With ML you don't have to create all the rules
- However, you have to manually annotate a training corpus (or get someone else to do it!)
- Rule-based approaches (e.g. JAPE) and ML work well together; JAPE is often used extensively to prepare data for ML

Terminology: Instances, attributes, classes

California Governor Arnold Schwarzenegger proposes deep cuts.

Instances: Any annotation

Tokens are often convenient

Token | Token | Token | Token | Token | Tok | Tok

Attributes: Any annotation feature relative to instances

Token.String

Token.category (POS)

Sentence.length

Sentence

Class: The thing we want to learn

A feature on an annotation

Entity.type =Location

Entity.type=Person

Instances

- Instances are cases that may be learned
- Every instance is a decision for the ML algorithm to make
- To which class does this instance belong?
  - "California"→Location

Attributes

- Attributes are pieces of information about instances
- They are sometimes called “features” in machine learning literature
- Examples
  - Token.string == “Arnold”
  - Token.orth == upperInitial
  - Token(-1).string == “Governor”

Classes

- The class is what we want to learn
- Suppose we want to find persons' names: for every instance, the question is “is this a person name?” and the classes are “yes” and “no”
- Sometimes there are many classes, for example we may want to learn entity types
  - For every instance, the question is “which type from the list does this instance belong to?”
  - One answer is “none of them”

ML Tasks

GATE supports 3 types of ML tasks:

- chunk recognition (named entity recognition, NP chunking)
- text classification (sentiment classification, POS tagging)
- relation annotation

Training

- Training involves presenting data to the ML algorithm from which it creates a model
- The training data (instances) have been annotated with class annotations as well as attributes
- Models are representations of decision-making processes that allow the machine learner to decide what class the instance has based on the attributes of the instance

Application

- When the machine learner is applied, it creates new class annotations on data using the model
- The corpus it is applied to must contain the required attribute annotations
- The machine learner will work best if the application data is similar to the training data

Evaluation

- We want to know how good our machine learner is before we use it for a real task
- Therefore we apply it to some data for which we already have class annotations
  - The “right answers”, sometimes called “gold standard”
- If the machine learner creates the same annotations as the gold standard, then we know it is performing well
- The test corpus must not be the same corpus as you trained on
  - This would give the machine learner an advantage, and would give a false idea of how good it is
- GATE's ML PR has a built-in evaluation mode that splits the corpus into training and test sets and cross-validates them

Setting up a Corpus

Load the corpus

- Create a corpus (any name is fine)
- Populate it from module-11-hands-on/corpus/*.xml in your hands-on materials
- Use UTF-8 encoding
- Open a document and examine its annotations

Examining the corpus

- The corpus contains an annotation set called “Key”, which has been manually prepared
- Within this annotation set are annotations of types “Date”, “Location”, “Money”, “Organization” and so forth
- There are also some annotations in the “Original markups” set

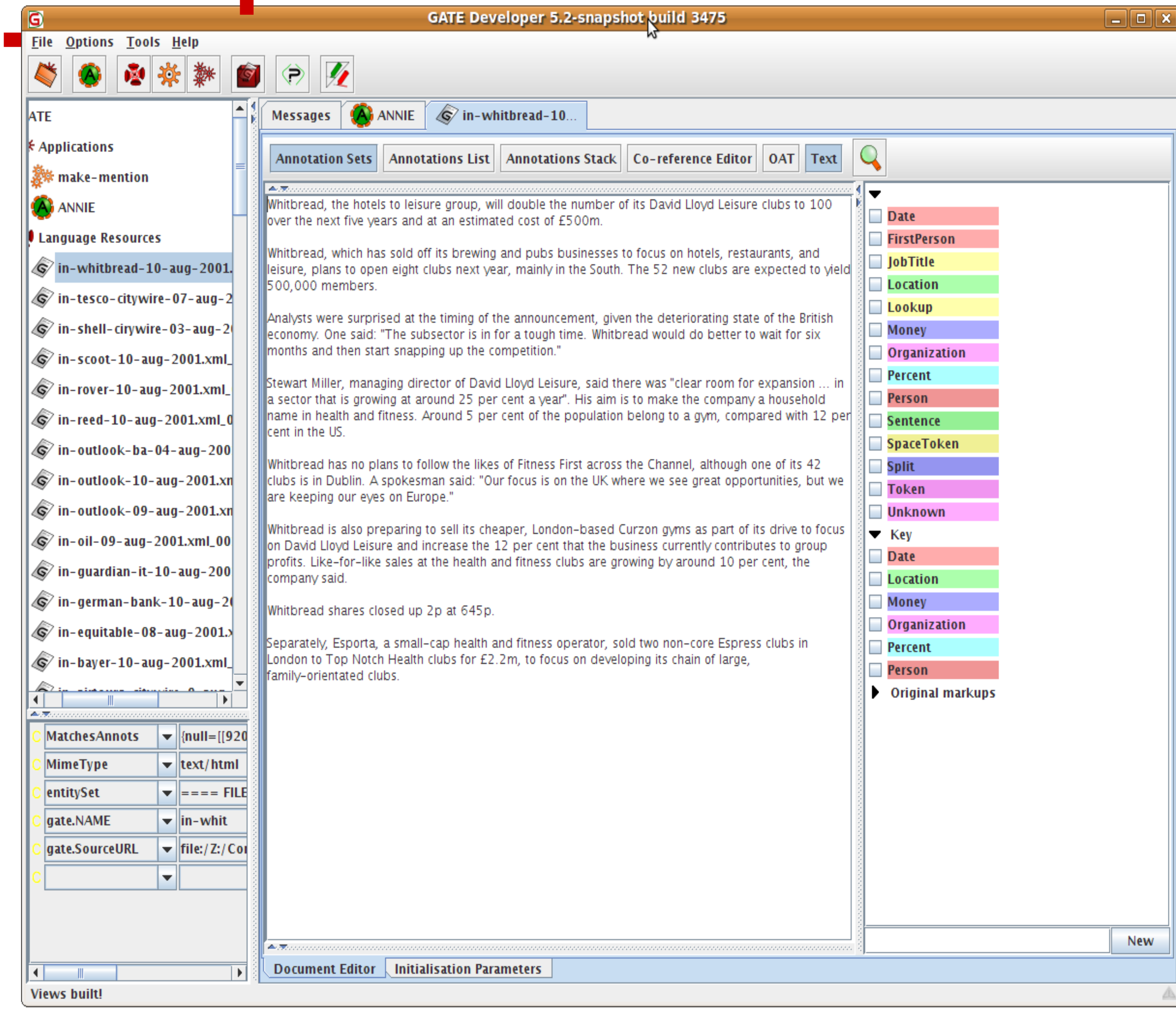
What are we going to use this corpus for?

- We are going to train a machine learner to annotate corpora with these entity types
- We need a training corpus and a test corpus
- The training corpus will be used by the machine learner to deduce relationships between attributes and entity types (classes)
- The test corpus will be used to find out how well it is working, by comparing annotations created by the learner with the correct annotations that are already there
- In *Evaluation* mode, which we will try first, the ML PR automatically splits one corpus up into training and test sets

Instances and Attributes

- This corpus so far contains only the class annotations
- There is not much in this corpus to learn from
- What would our instances be?
- What would our attributes be?
- If we run ANNIE over the corpus, then we can use “Token” annotations for instances, and we would have various options for attributes
- **Load ANNIE but add the Key AS to setsToKeep in the document reset PR!**
- **Run ANNIE over your corpus**

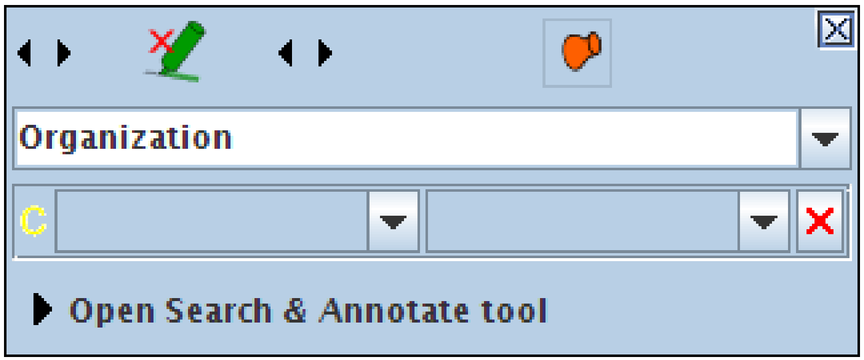
Running ANNIE on the corpus



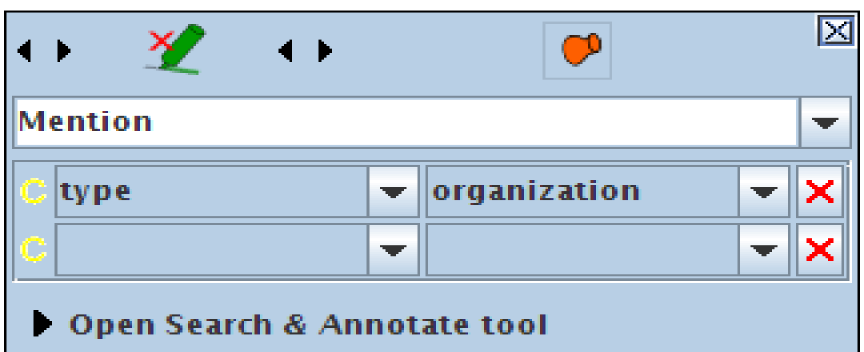
- Having run ANNIE on the corpus, we have more annotations to work with

Preparing the corpus: Classes

- What we have:



- What we need:



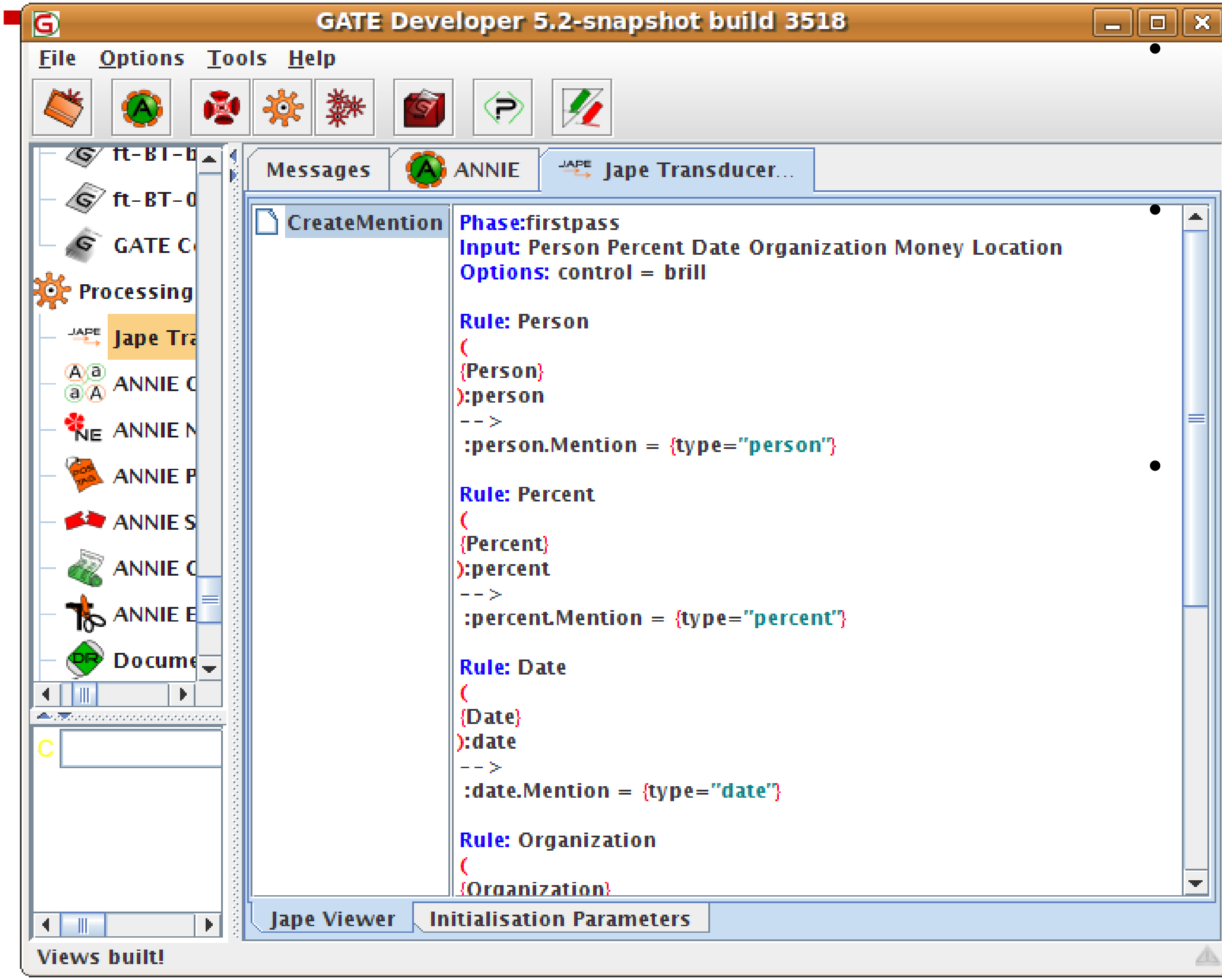
Preparing the corpus: Classes

- Currently each class has its own annotation type (Date, Person, Percent etc.)
- But the ML PR expects the class to be a feature value, not an annotation type
- Therefore we are going to make a new annotation type for the ML to learn from, “Mention” (it doesn't matter what it's called as long as we're consistent)

Making class annotations

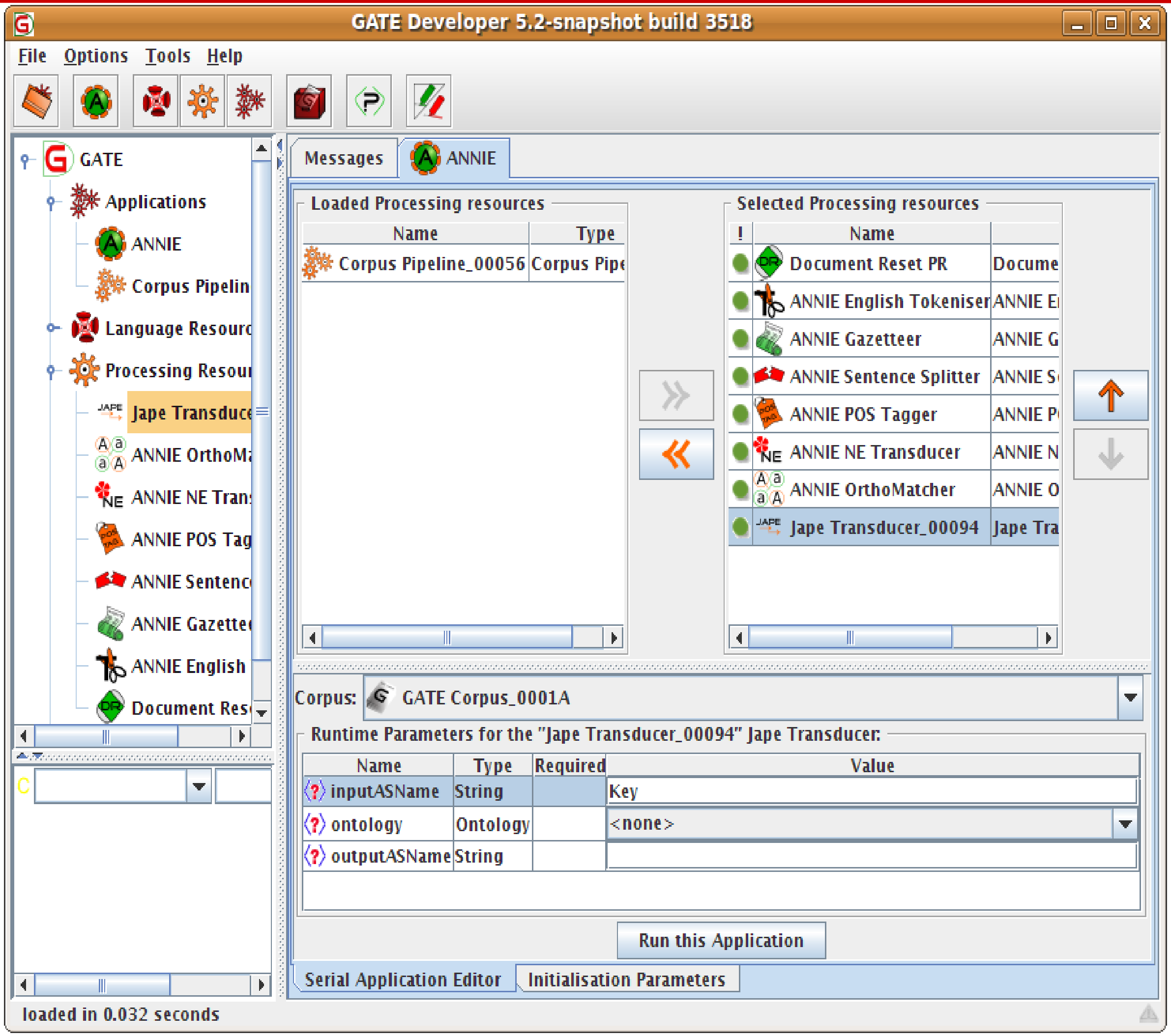
- Load a JAPE transducer from the module-11-hands-on/CreateMention.jape grammar
- Look at the grammar in GATE

The CreateMention.jape grammar



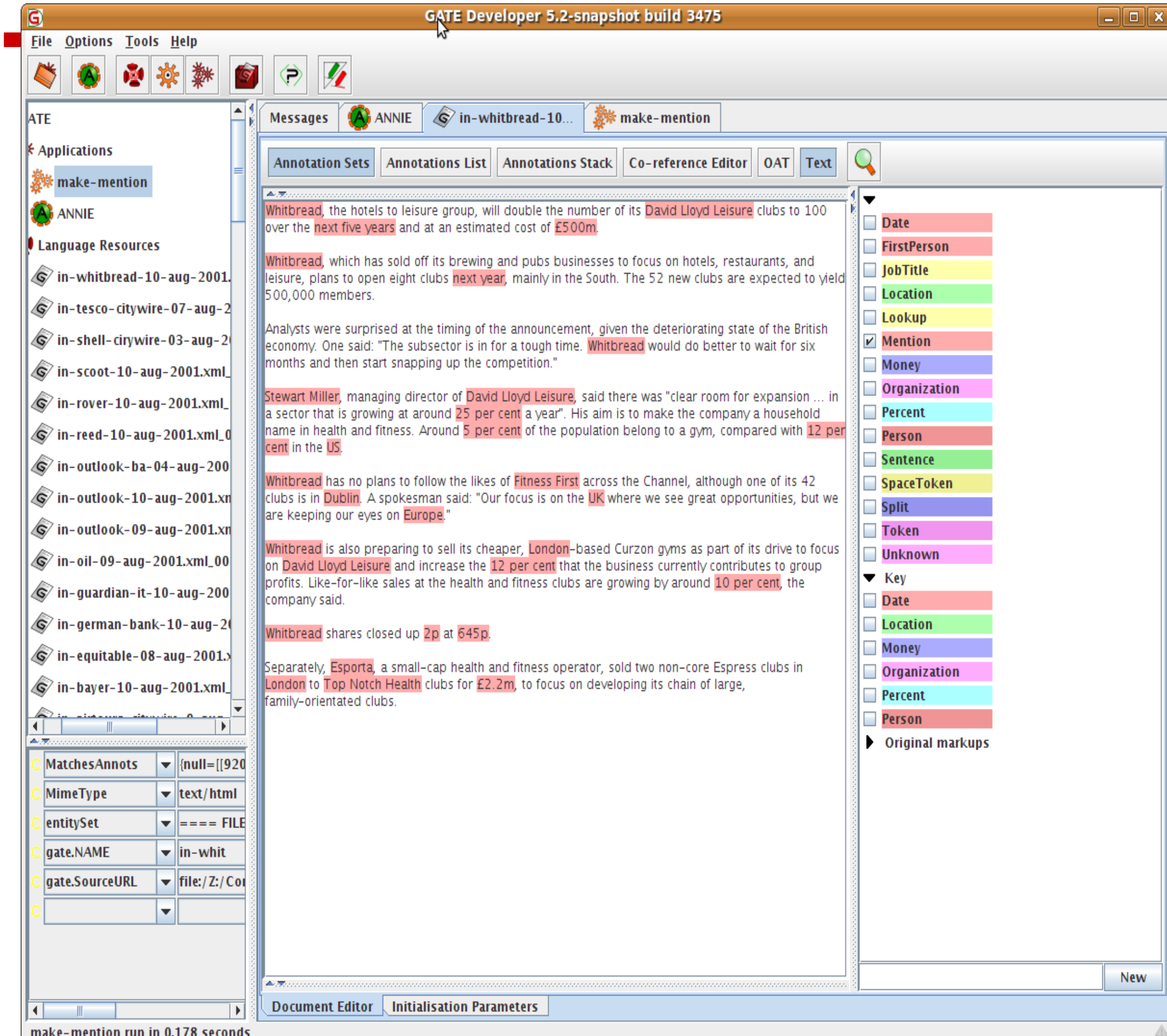
- This grammar makes a new annotation type called "Mention"
- It makes the previous annotation type into a feature of the "Mention" annotation
- Feature name is "type" because "class" is reserved for ontologies

Applying the grammar to the corpus



- Add the JAPE transducer at the end of your ANNIE application
- Set the inputASName to “Key”
- Leave the outputASName blank (default)

Check the "Mention" annotations



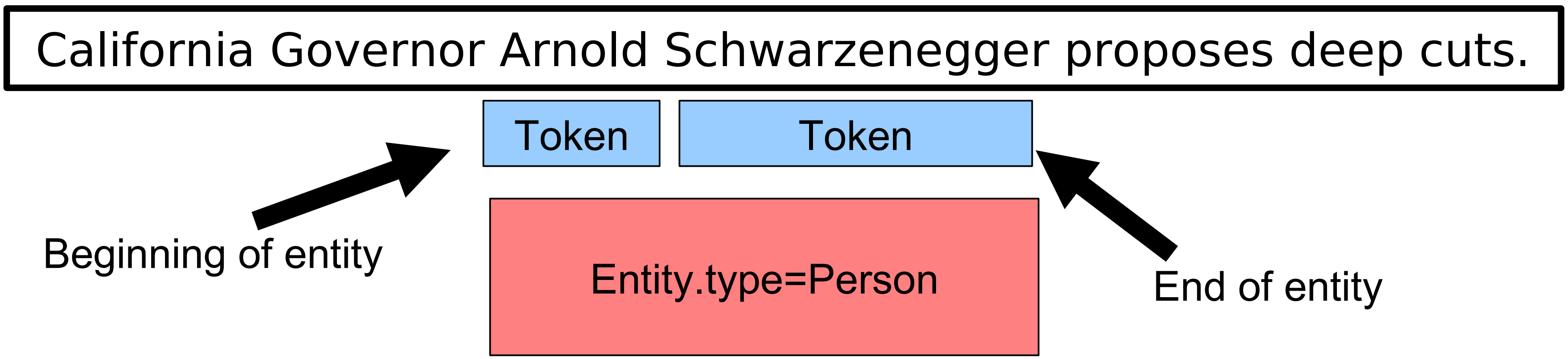
- Rerun the application
- Check that you have some "Mention" annotations
- Check that they have a feature "type" and that the values look right

The Configuration File

Looking at the configuration file

- In the configuration file, we tell the machine learning PR what we want it to do
- You will find a configuration file in your hands-on materials, called ml-config-file.xml
- **Open it using a text editor**

<SURROUND value="true"/>



- This class to be learned covers more than one instance; the PR has to learn the boundaries
- So ***surround mode***
- Transparent to the user

Confidence Thresholds

```
<PARAMETER name="thresholdProbabilityEntity" value="0.2"/>
<PARAMETER name="thresholdProbabilityBoundary" value="0.4"/>
```

- Classifiers provide confidence ratings—how likely a result is to be correct
- We must determine how certain is good enough
- Depending on the application we might prefer to include or exclude annotations for which the learner is not too sure
- thresholdProbabilityBoundary is a threshold for the beginning and end instances
- thresholdProbabilityEntity is a threshold for beginning and end instances combined

<multiClassification2Binary method="one-vs-others"/>

California Governor Arnold Schwarzenegger proposes deep cuts.

Entity.type =Location

Entity.type=Person

- Many algorithms are binary classifiers (e.g. yes/no)
- We have several classes (Person, Location, Organization etc.)
- Therefore the problem must be converted to a set of binary problems, so we can use binary algorithms
- **one-vs-others**
  - LOC vs PERS+ORG / PERS vs LOC+ORG / ORG vs LOC+PERS
- **one-vs-another**
  - LOC vs PERS / LOC vs ORG / PERS vs ORG

<multiClassification2Binary method="one-vs-others"/>

- With more than a few classes, **one-vs-another** becomes very computationally expensive!
- **one-vs-others**: N classes => N classifiers
  - A vs B+C+D, B vs A+C+D, C vs A+B+D, D vs A+B+C
- **one-vs-another**: N classes => N×(N-1)/2 classifiers
  - A vs B, A vs C, A vs D, B vs C, B vs D, C vs D

<EVALUATION method="holdout" ratio="0.66"/>

- We are going to evaluate our application in two ways today
  - The ML PR can automatically evaluate for us
  - We will also run our own evaluation
- This parameter dictates how the ML PR will evaluate for us, if we run it in evaluation mode
- We are telling it that it should reserve a third of the data as a test set, train, then apply the result to the held out set
- Alternatively, we could ask the PR to run a cross-validation evaluation

Evaluation

```
<EVALUATION method="kfold" runs="10"/>
```

OR

```
<EVALUATION method="holdout" ratio="0.66"/>
```

- Holdout randomly picks *ratio* documents for training and uses the rest for testing; this is faster than k-fold because it only runs once
- But k-fold cross-validation will give you more reliable results and lets you "stretch" your corpus

K-Fold Cross-Validation

- In k-fold cross-validation, the corpus is split into k equal parts, and the learner is trained k times on k-1 parts and evaluated on 1; the results are averaged
- For example, if k=4, the documents are split into groups A, B, C, & D, then:
  - train on A+B+C, test on D
  - train on A+B+D, test on C
  - train on A+C+D, test on B
  - train on B+C+D, test on A
  - average these 4 results
- This maximises the use of the training data without losing testing accuracy, but takes 4 times as long
- <EVALUATION method="kfold" runs="4"/>

<ENGINE nickname="PAUM" ..

- Next we specify what machine learning algorithm we wish to use
- Today we are using the perceptron with uneven margins (“PAUM”)
- We will use the following options: options="-p 50 -n 5 -optB 0.3"
  - Challenge: find out what these options do! (Hint: user guide §17.2)

<INSTANCE-TYPE>Token</INSTANCE-TYPE>

- Next, we tell the ML PR what our instance annotation is
- The goal of the ML PR is to try to learn how the attributes of every instance relate to its class, so the instance is an important choice
- We have decided that the “Token” is our instance annotation type
  - We made sure, earlier, that we have “Token” annotations in our corpus

Specifying Attributes

```
<ATTRIBUTELIST>
  <NAME>Form</NAME>
  <SEMTYPE>NOMINAL</SEMTYPE>
  <TYPE>Token</TYPE>
  <FEATURE>category</FEATURE>
  <RANGE from="-2" to="2"/>
</ATTRIBUTELIST>
```

- For every attribute, we create a specification like the one above
- This is the information from which the PR will learn, so it is important to give it some good data
- You can see in the configuration file that there are several attributes, providing a good range of information
- However, if you have too many attributes it can take a very long time to learn!

Breaking down the attribute specification

- <NAME>Form</NAME>
  - This is the name that we choose for this attribute. It can be anything we want, but it will help us later if we make it something sensible!
- <SEMTYPE>NOMINAL</SEMTYPE>
  - Is the value of this attribute a number or a name?

Breaking down the attribute specification

- <TYPE>Token</TYPE>
  - The value of the attribute will be taken from the “Token” annotation
- <FEATURE>category</FEATURE>
  - The value of the attribute will be taken from the “category” feature

Breaking down the attribute specification

```
<ATTRIBUTELIST>
  :
  <RANGE from="-2" to="2"/>
</ATTRIBUTELIST>
```

- Because this is an “ATTRIBUTELIST” specification, we can specify a “RANGE”
- In this case, we will gather attributes from the current instance and also the preceding and following two

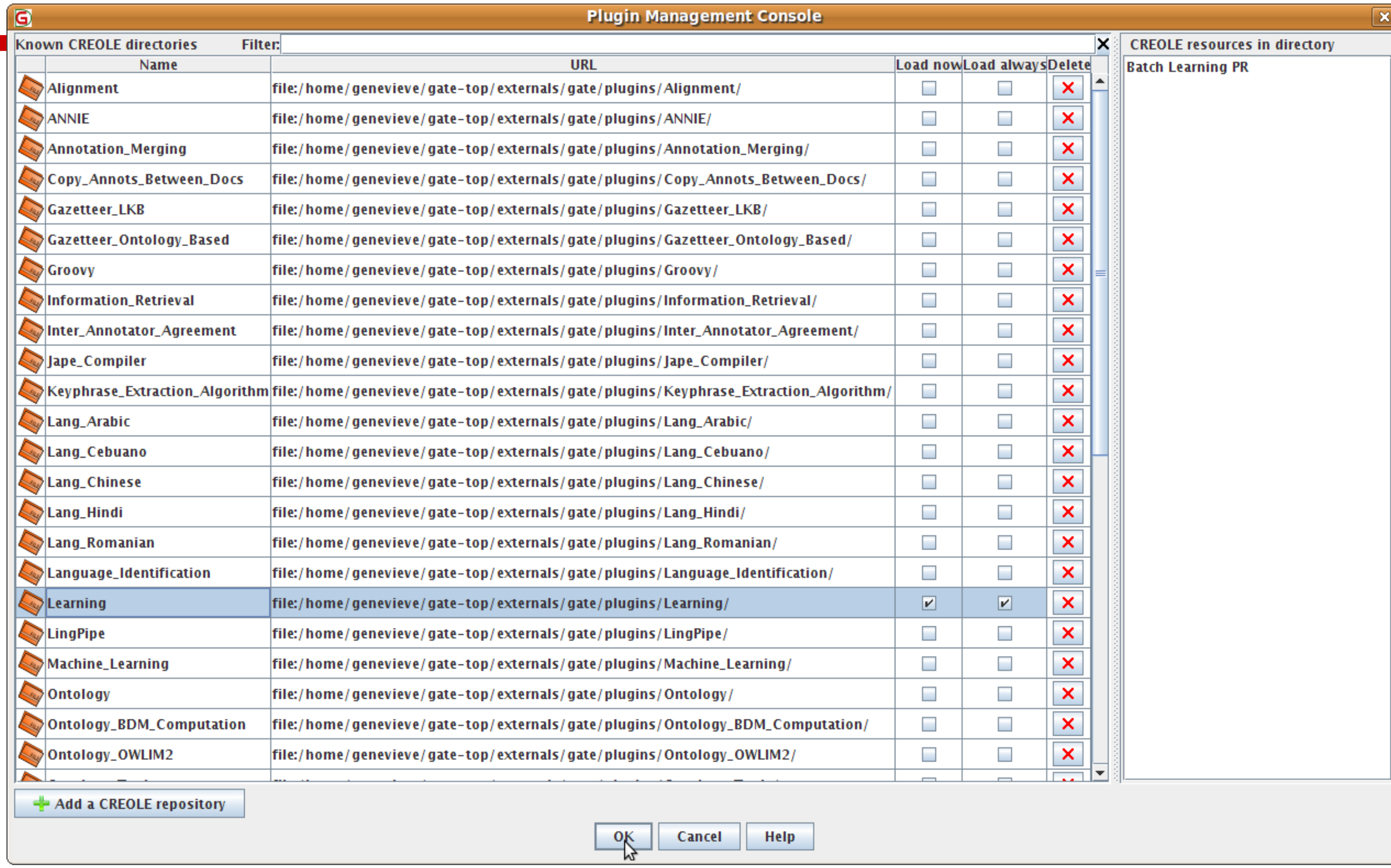
Specifying the Class Attribute

```
<ATTRIBUTE>
    <NAME>Class</NAME>
    <SEMTYPE>NOMINAL</SEMTYPE>
    <TYPE>Mention</TYPE>
    <FEATURE>type</FEATURE>
    <POSITION>0</POSITION>
    <CLASS/>
</ATTRIBUTE>
```

- You can call the class attribute whatever you want, but “Class” is a sensible choice
- Remember that our class attribute is the “type” feature of the “Mention” annotation
- This is an ATTRIBUTE, not an ATTRIBUTELIST, so we have “position”, not “range”
- The <CLASS/> element tells the Batch Learning PR that this is the class attribute to learn.

Running the ML PR in evaluation mode

Loading the Learning plugin

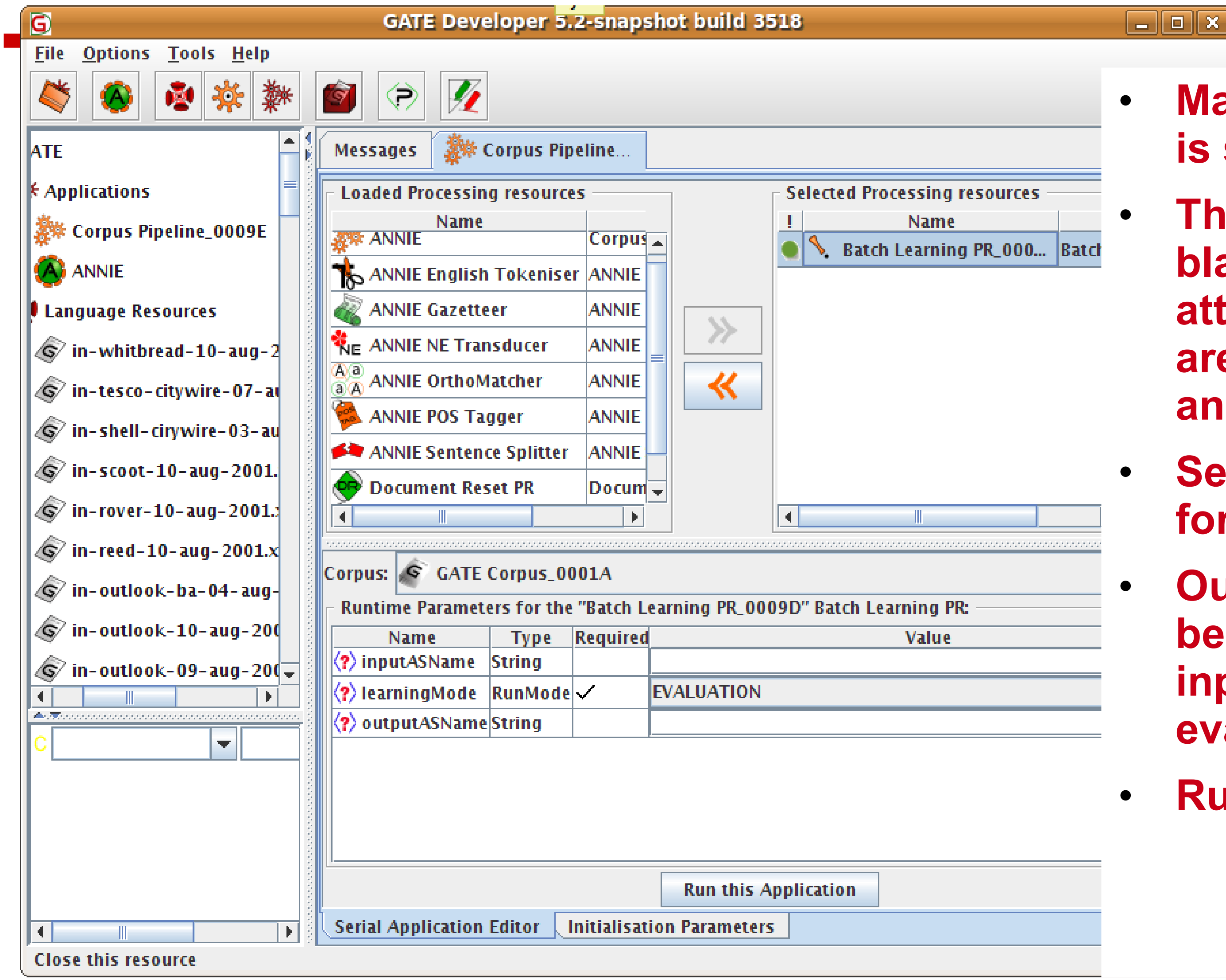


- **Load the "Learning" plugin**
- (We are **not** going to use the "Machine Learning" plugin, which is obsolete and does not have all the functionality we want.)

Creating a learning application

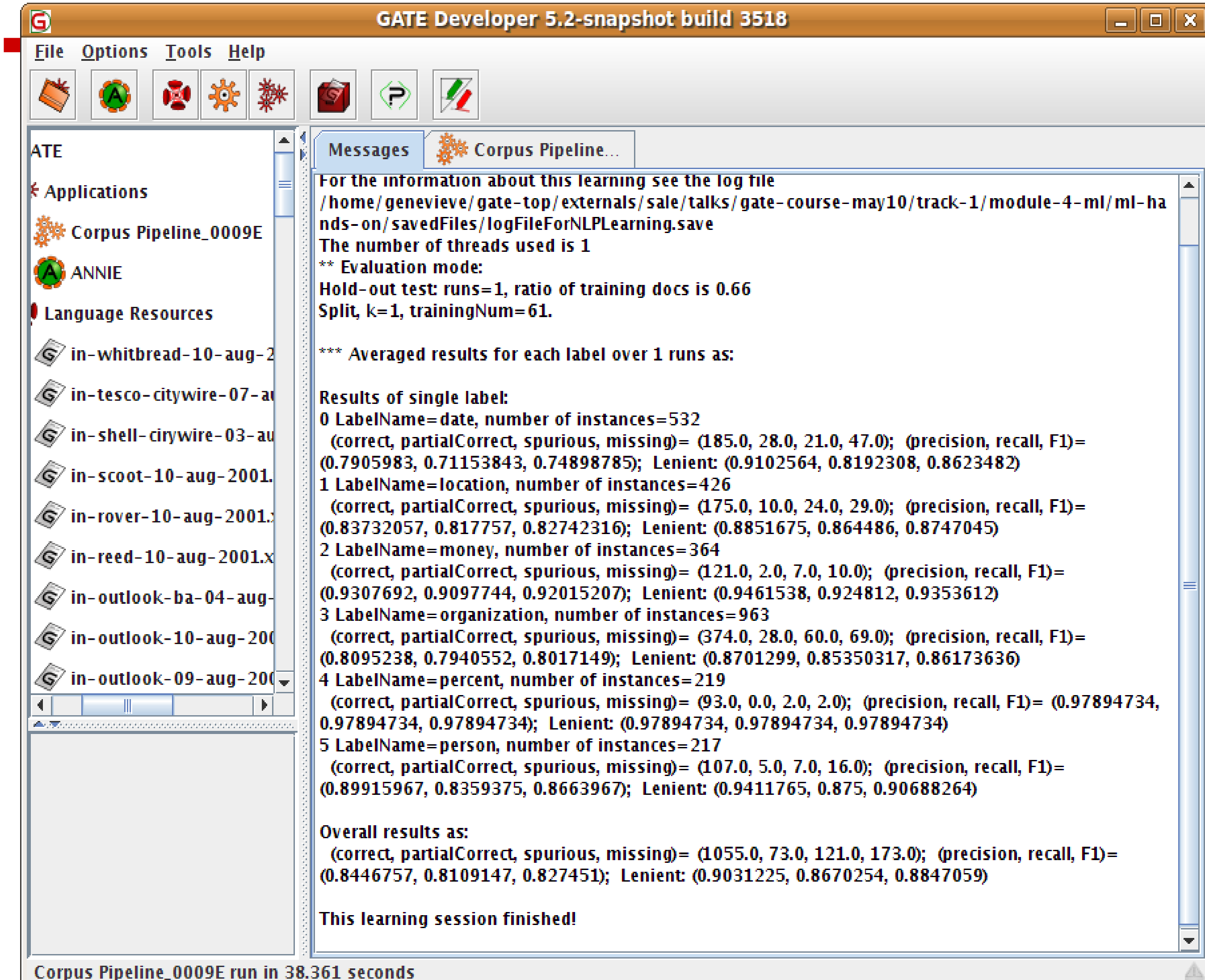
- Create a “Batch Learning PR” using your configuration file
- Make a new corpus pipeline and put this PR in it

Running the application in evaluation mode



- **Make sure the corpus is selected**
- **The inputASName is blank because the attributes and classes are in the default annotation set**
- **Select “EVALUATION” for the learningMode**
- **OutputASName should be the same as inputASName in evaluation mode**
- **Run the application!**

Inspecting the results



- The application may take a few minutes to run
- **When it is finished, switch to the “Messages” tab to examine the results**

How well did we do?

- Here is my result:

(precision, recall, F1)= (0.8462151, 0.81629515, 0.83098596)

- These figures look pretty good, but what do they mean?
- Next we will discuss evaluation measures
- Then we will run the PR in different modes
- Then we will see if we can get these numbers any higher!

Evaluation in Machine Learning

Recap of Evaluation in GATE

- Evaluation is an important part of information extraction work
  - We need to find out how good our application is by comparing its annotations to the “right answers” (manually prepared or corrected annotations)
  - Sometimes we need to compare annotations by different annotators, to see how consistent they are
- We use similar functions for both types of evaluation tasks

Evaluation Mode

- We ran the machine learning PR in evaluation mode earlier
- We specified how the PR should run evaluation in the configuration file
- Once we had run the application, we obtained evaluation statistics in the “Messages” tab

What are precision, recall and F1?

- Precision: what proportion of our ML annotations were correct?
- Recall: what proportion of the correct annotations did our ML create?
- P = correct / (correct + spurious) = tp / (tp + fp)
- R = correct / (correct + missing) = tp / (tp + fn)
- where tp = true positives, fp = false positives, fn = false negatives

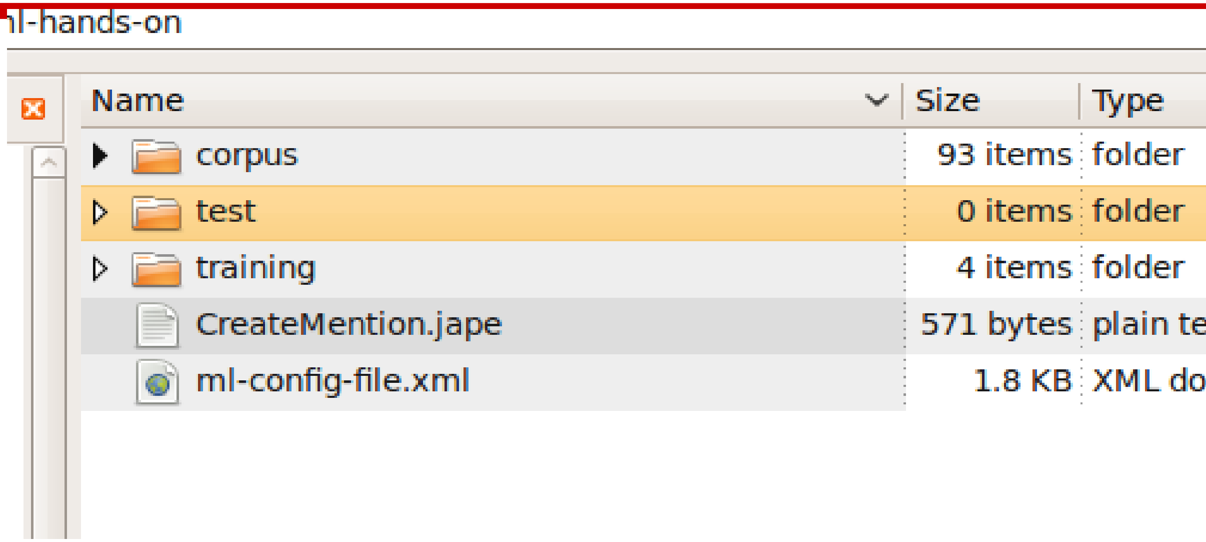
What are precision, recall and F1?

- F-score is an amalgam of the two measures
  - $F = 1 / (\beta/P + (1-\beta)/R)$
  - $F1 = 2PR / (R + P)$
  - The equally balanced F1 ($\beta = 0.5$) is the most common F-measure
- We can also run our own ML evaluation using the Corpus QA tool—let's do that now

Splitting into training and test corpora

- As mentioned earlier, to truly know how well a machine learner is performing, you need to test it on data that it was not trained on
- We need separate test and training corpora
- So now we are going to split our corpus in two

Saving and splitting the corpus

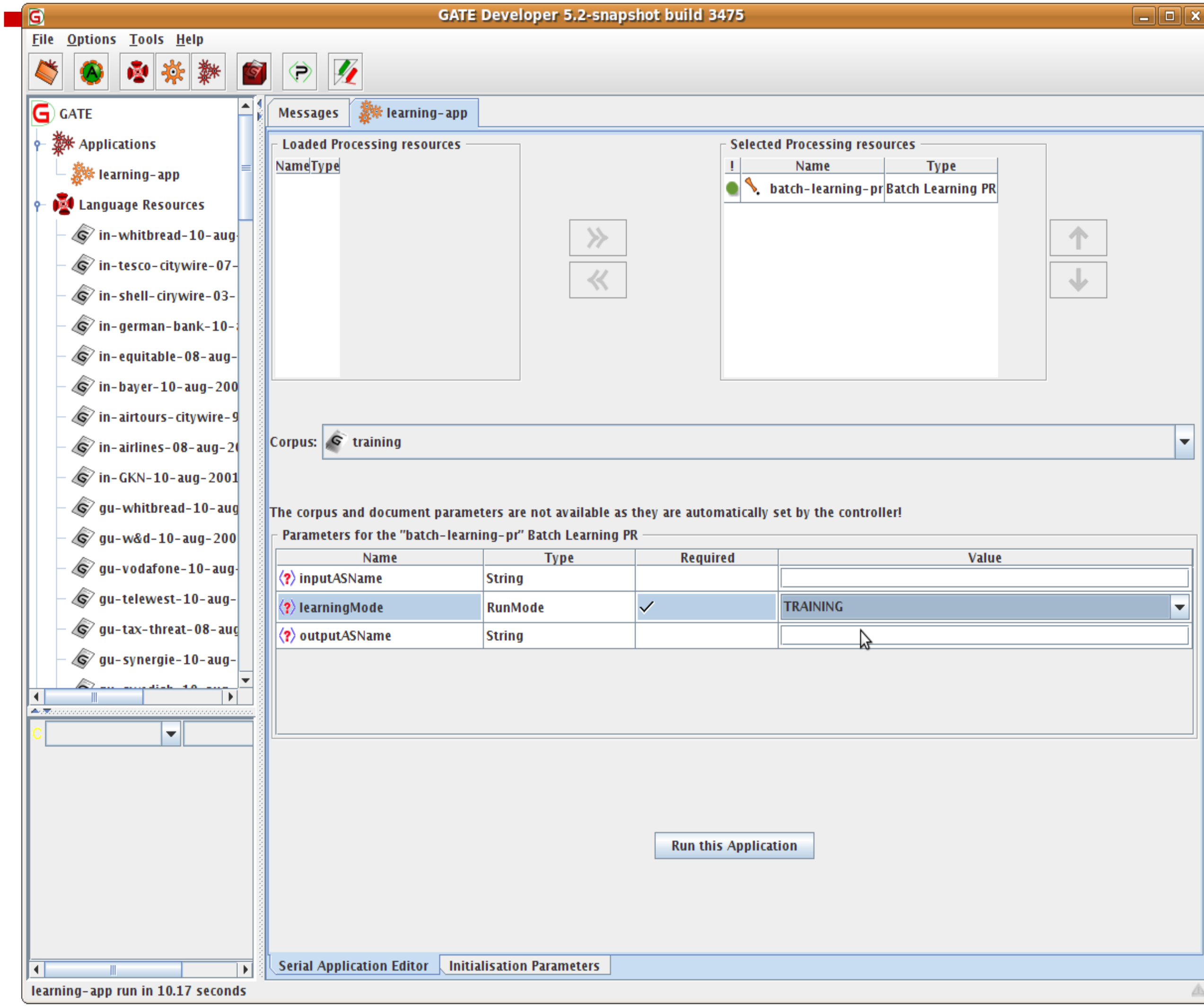


- Right click on your corpus and select “Save as XML”
- Create a new folder called “training” and save the documents in it
- Use your file manager to create a new directory alongside it called “test”
- Use your file manager to pick half the documents in “training” and move them into “test” (try to randomise them a bit)

Tidying up

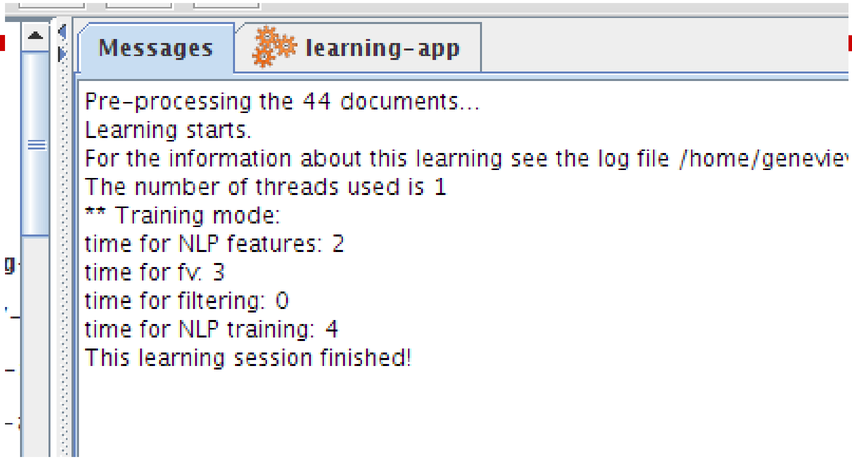
- Close all your open documents and corpora in GATE Developer
- Close the modified ANNIE application recursively
- Create new corpora called “training” and “test”
- Populate your corpora with the documents you saved to disk
  - As before, use UTF-8

Running the ML PR in Training Mode



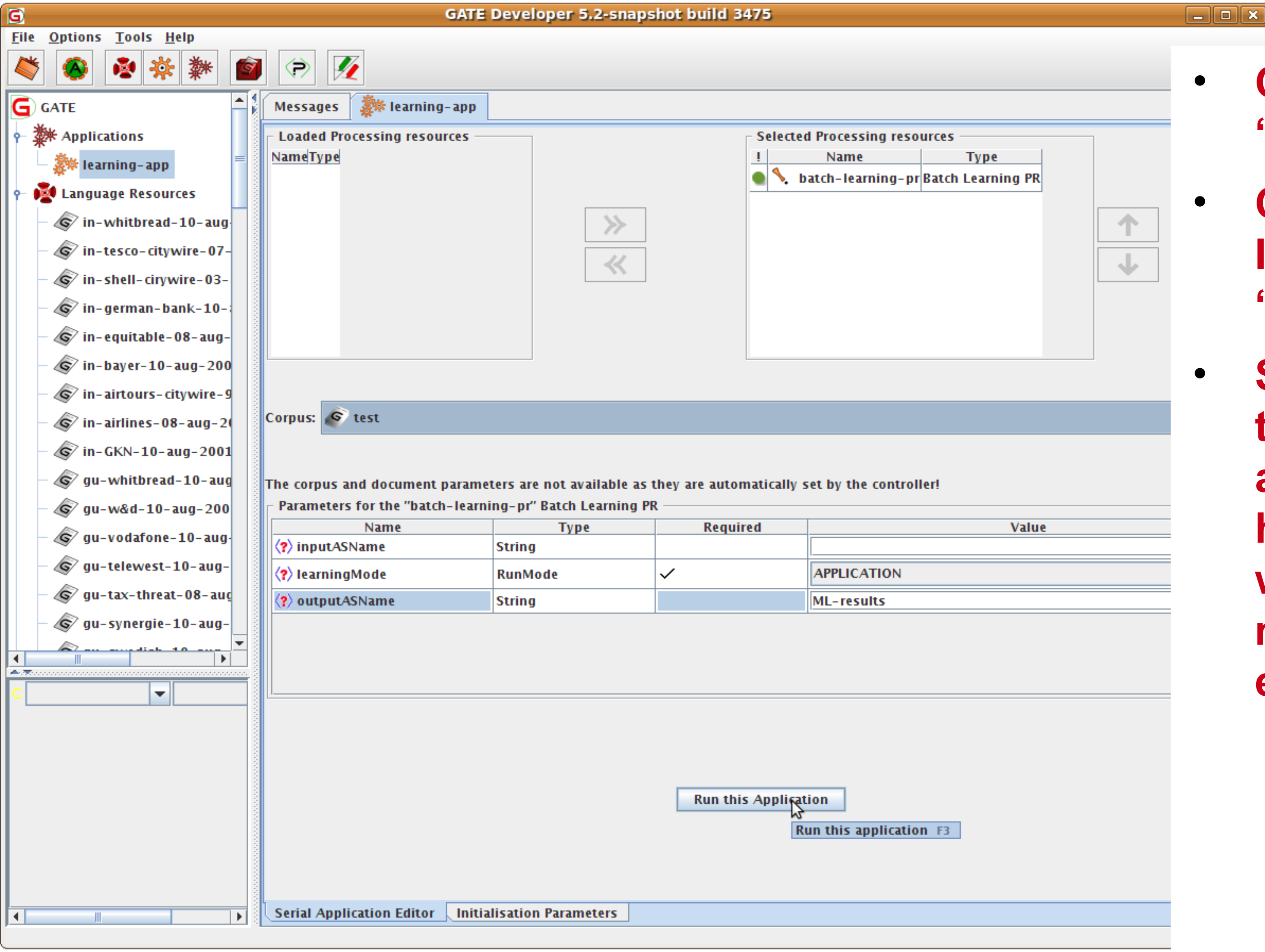
- Check that your PR is set to run on the training corpus
- Change the learningMode to “TRAINING” (the outputASName doesn't matter)
- Run the application

Finished Training!



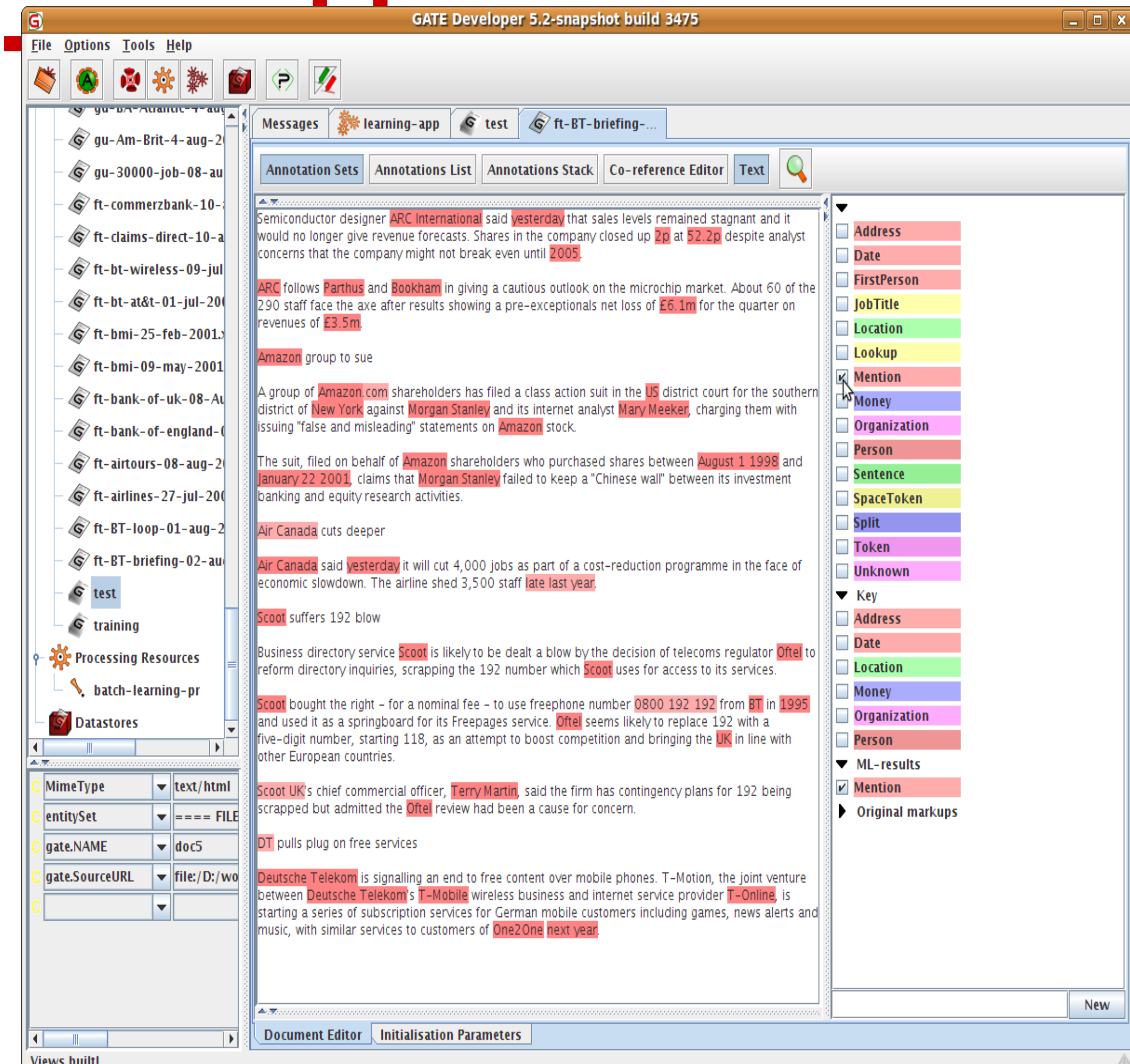
- Training may take a few minutes
- This time there is no evaluation result in the messages tab

Running the ML PR in Application Mode



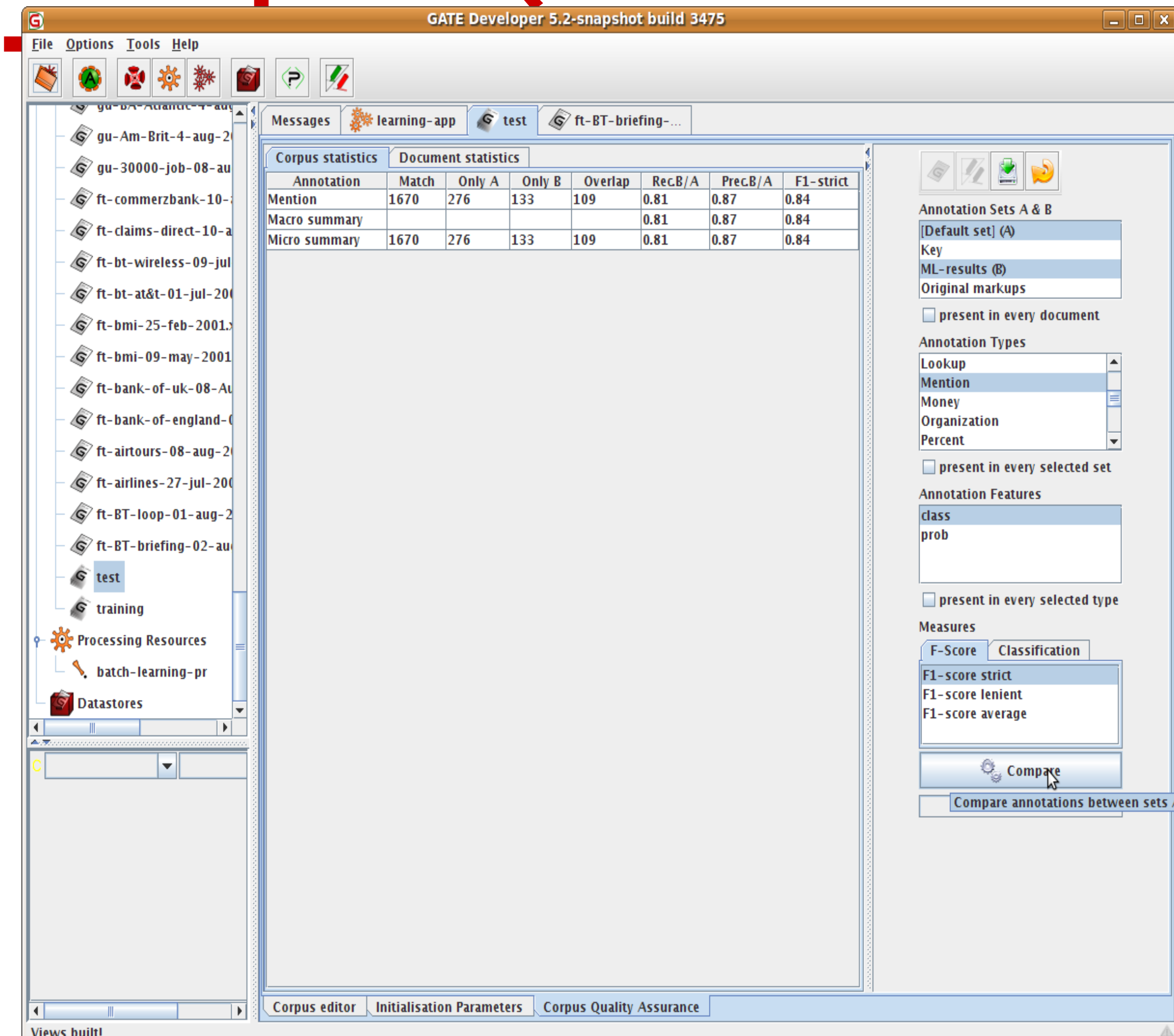
- Change corpus to "test"
- Change learningMode to "APPLICATION"
- Set outputASName to "ML": your new annotations will go here, and you don't want to get them mixed up with the existing ones!

Examining the results of application



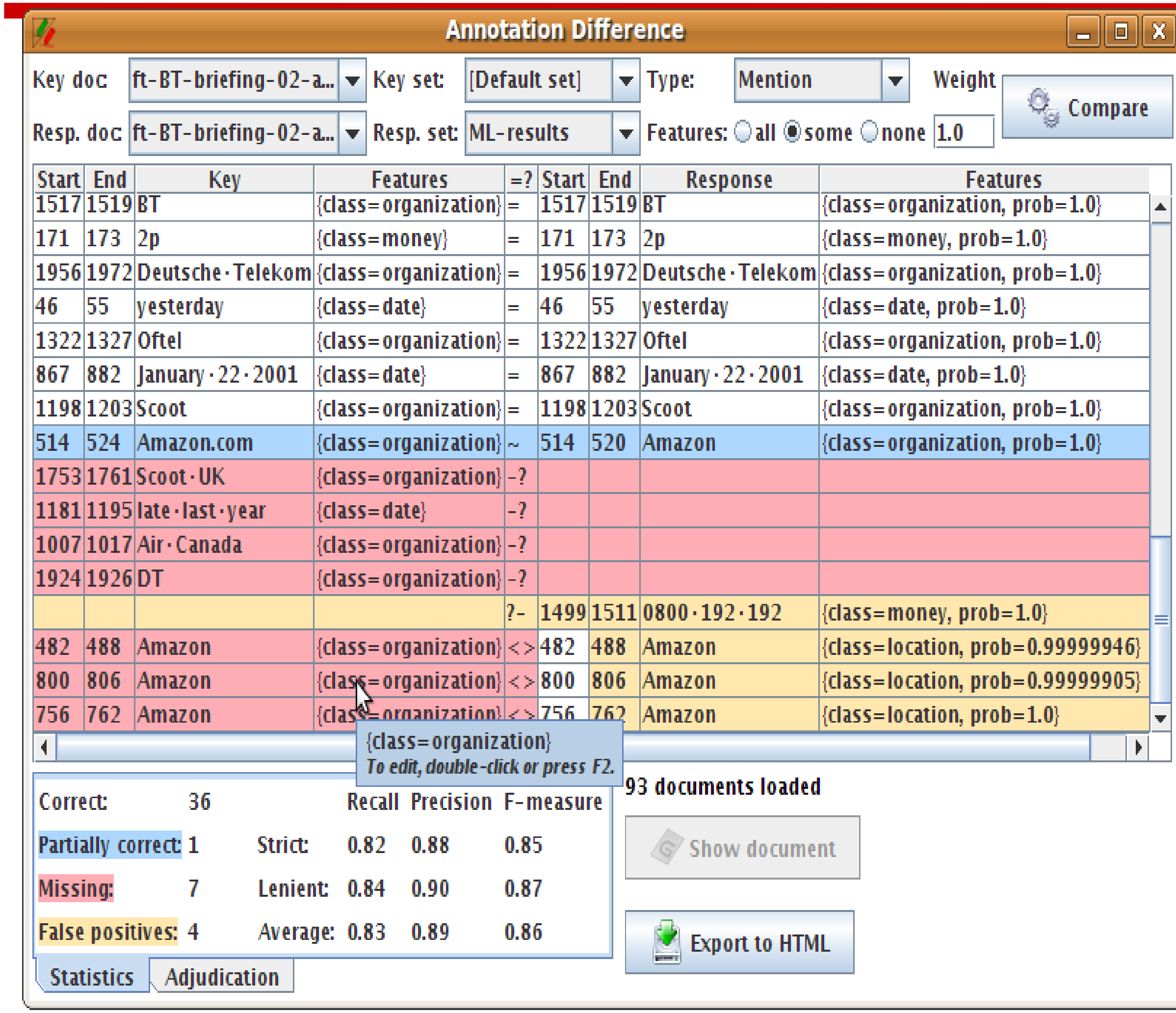
- **Choose a document from the test corpus to look at**
- You should have a new annotation set, created by the ML application
- There will be a “Mention” type both in the new set and the original
- They are similar but not identical!
- How similar do they appear to be? Do you think you will get a good result?

Comparing the Sets with Corpus QA



- Select the test corpus and click on the Corpus QA tab (it will take a few seconds to scan the document)
- Select the Default and ML annotation sets
- Select the “Mention” type
- Select the “type” feature
- Choose an F-measure
- Click on Compare
- Did you get a good result? How does it compare to the result you got using evaluation mode?

Using Annotation Diff to examine performance



- **Switch to the “Document statistics” tab**
- **Choose a document**
- **Click on the Annotation Diff icon**
- What kind of mistakes did your application make?

Varying the configuration file

- Now we are going to experiment with varying the configuration file to see if we can produce varied results
- You can edit the configuration file in your text editor
- Make sure you save your changes then **reinitialise the PR** (this reads the file again)

Exercises

- Spend some time working on your exercise sheet
- Feel free to ask questions

Confidence Thresholds

```
<PARAMETER name="thresholdProbabilityEntity" value="0.2"/>
<PARAMETER name="thresholdProbabilityBoundary" value="0.42"/>
<PARAMETER name="thresholdProbabilityClassification" value="0.5"/>
```

- Each classifier will provide confidence ratings—how likely is a result to be correct; we must determine how certain is good enough
- Depending on the application we might prefer to include or exclude annotations for which the learner is not too sure
- thresholdProbabilityBoundary and thresholdProbabilityEntity are thresholds for chunk learning
- thresholdProbabilityClassification applies to classification tasks, such as relation learning

Classification tasks

- Example: the documents contains spans of text, which you want to classify as positive, negative, or neutral.
- This will be covered in more detail in Module 12 (Opinion Mining) tomorrow, with hands-on work

Classification tasks

- thresholdProbabilityClassification: the “pickiness” of the classifiers
  - increasing this generally raises precision and reduces recall
  - decreasing this generally increases recall and reduces precision
- thresholdProbabilityBoundary and thresholdProbabilityEntity: ignored

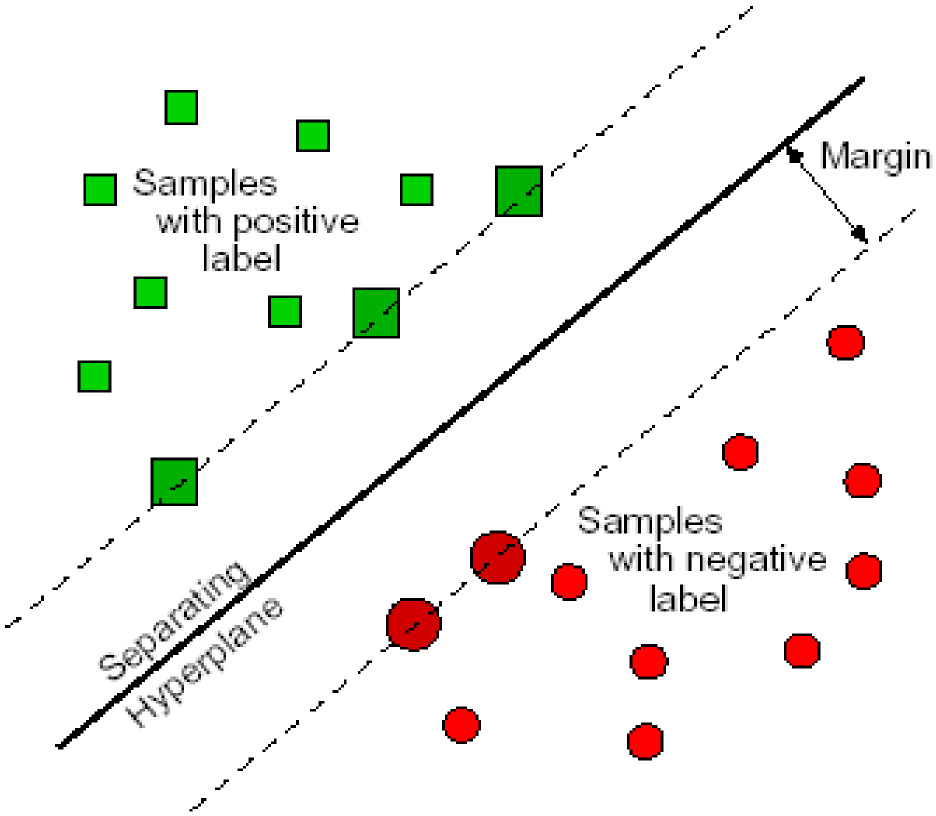
Classification tasks

- <SURROUND VALUE=”FALSE”/>
- INSTANCE-TYPE: type of annotation that covers each span of text to classify
- Typically use NGRAM elements as attributes
- The GATE user guide gives examples

Engines and Algorithms

Support Vector Machines

- Attempt to find a hyperplane that separates data
- Goal: maximize margin separating two classes
- Wider margin = greater generalisation

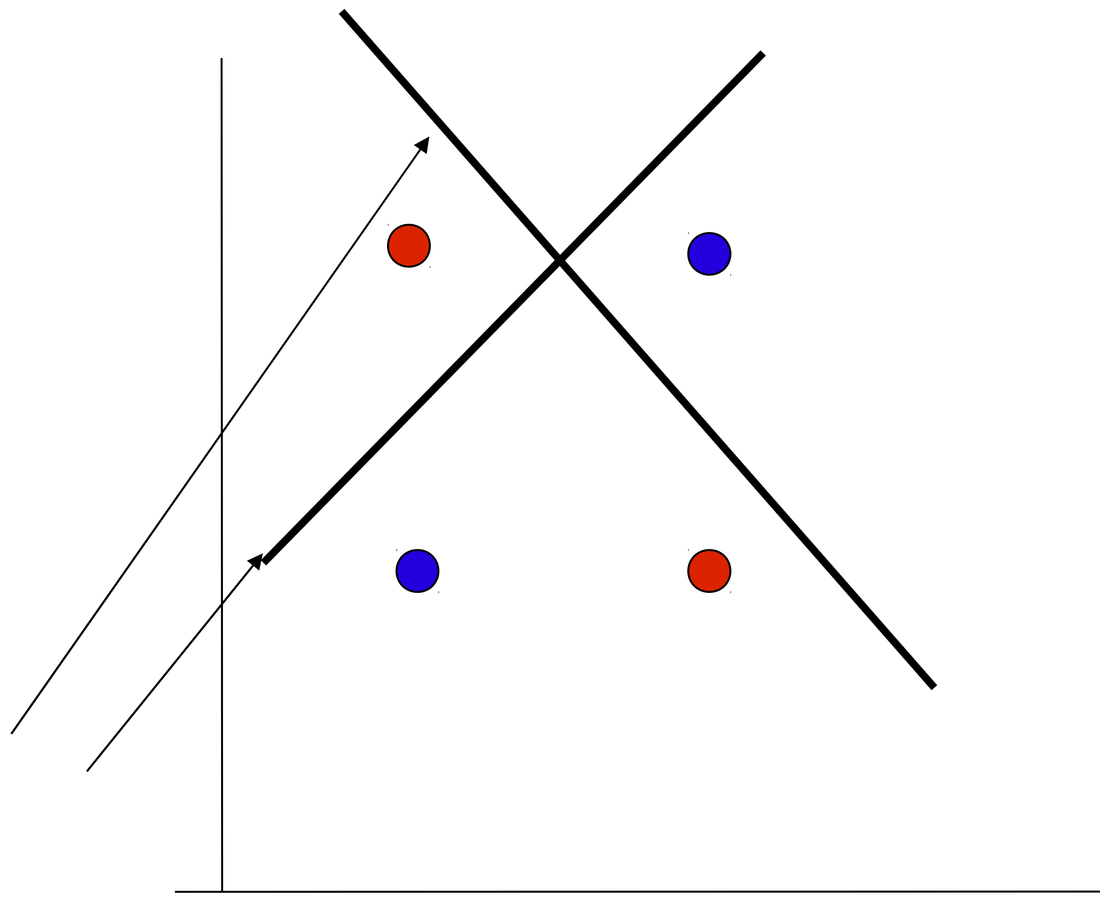


Support Vector Machines

- Points near decision boundary: support vectors (removing them would change boundary)
- Points far from boundary not important for decision
- What if data doesn't split?
  - Soft boundary methods exist for imperfect solutions
  - However linear separator may be completely unsuitable

Support Vector Machines

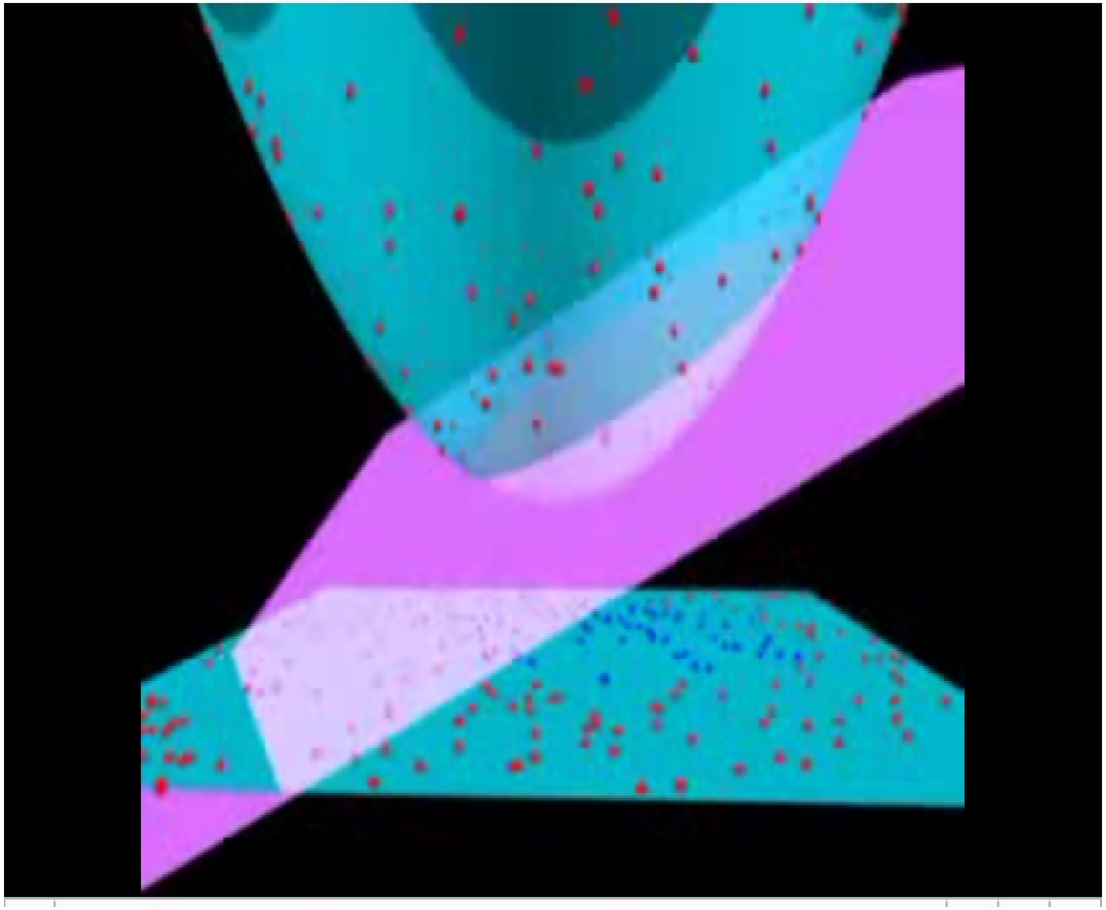
- What if there is no separating hyperplane?
- See example:
- Or class may be a globule



They do not work!

Kernel Trick

- Map data into different dimensionality
- http://www.youtube.com
- As shown in the video, due to polynomial kernel elliptical separators can be created nevertheless.
- Now the points are separable!



Kernel Trick in GATE and NLP

- Binomial kernel allows curved and elliptical separators to be created
- These are commonly used in language processing and are found to be successful
- Linear and polynomial kernels are implemented in Batch Learning PR's SVM

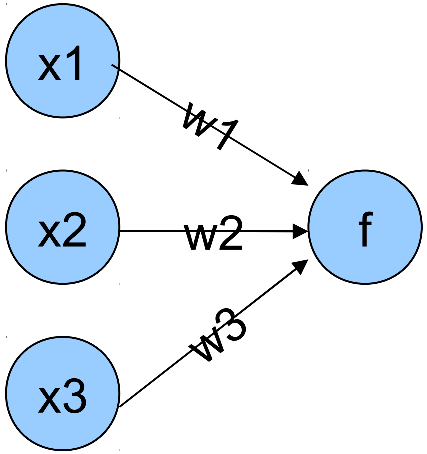
Support Vector Machines

- SVMs combined with kernel trick provide a powerful technique
- Multiclass methods simple extension to two class technique (one vs. another, one vs. others)
- Widely used with great success across a range of linguistic tasks

Perceptron and PAUM

- Perceptron is one of the oldest ML methods (invented in the 50s!)
- Has some similarities to SVM (it determines a hyperplane separator)
- Theoretically SVM works a little better because it calculates the optimal separator
- In practice, however, there is usually little difference, and Perceptron is a lot faster!

Perceptron



- You might think of perceptrons as being these things (correct)
- What this is actually calculating is a dot product w.x

More perceptron

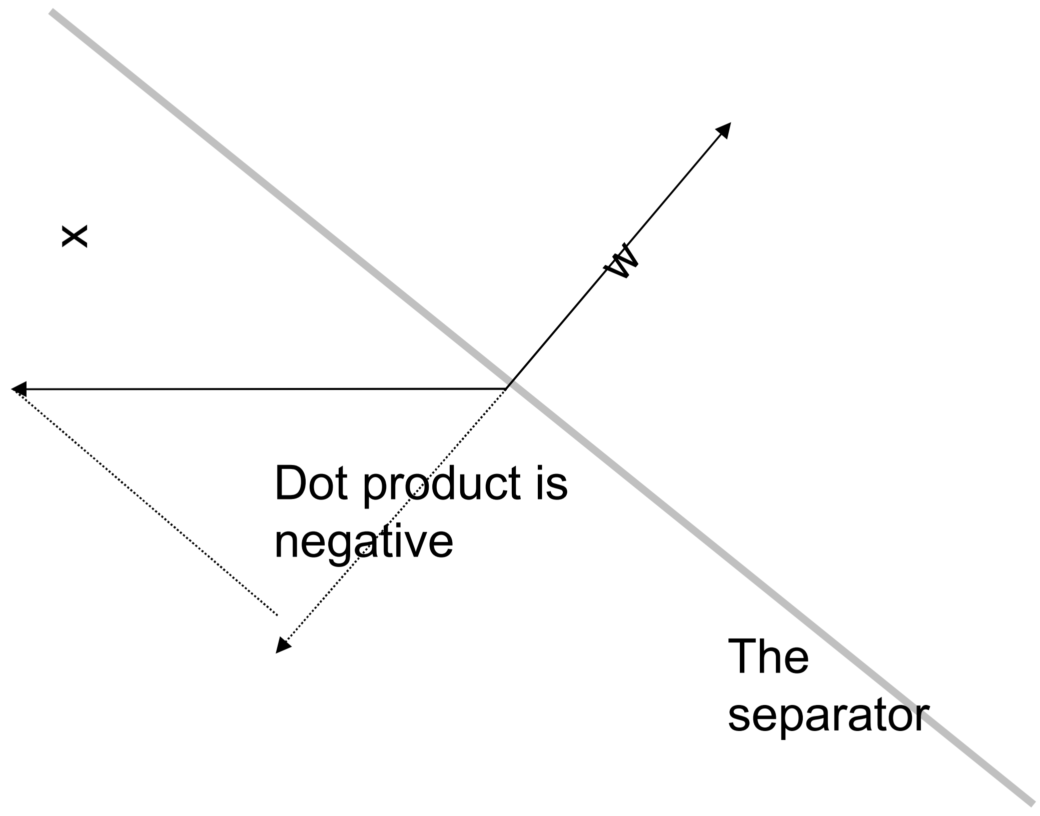
$$f(x) = \begin{cases} 1 & \text{if } \mathbf{w.x} + b > 0 \\ 0 & \text{otherwise} \end{cases}$$

- x is a datapoint represented as a vector
- w is a vector that defines the separating hyperplane (it is perpendicular to it)
- This function tells you which side of the hyperplane your point lies
- b defines an offset from the origin

More perceptron

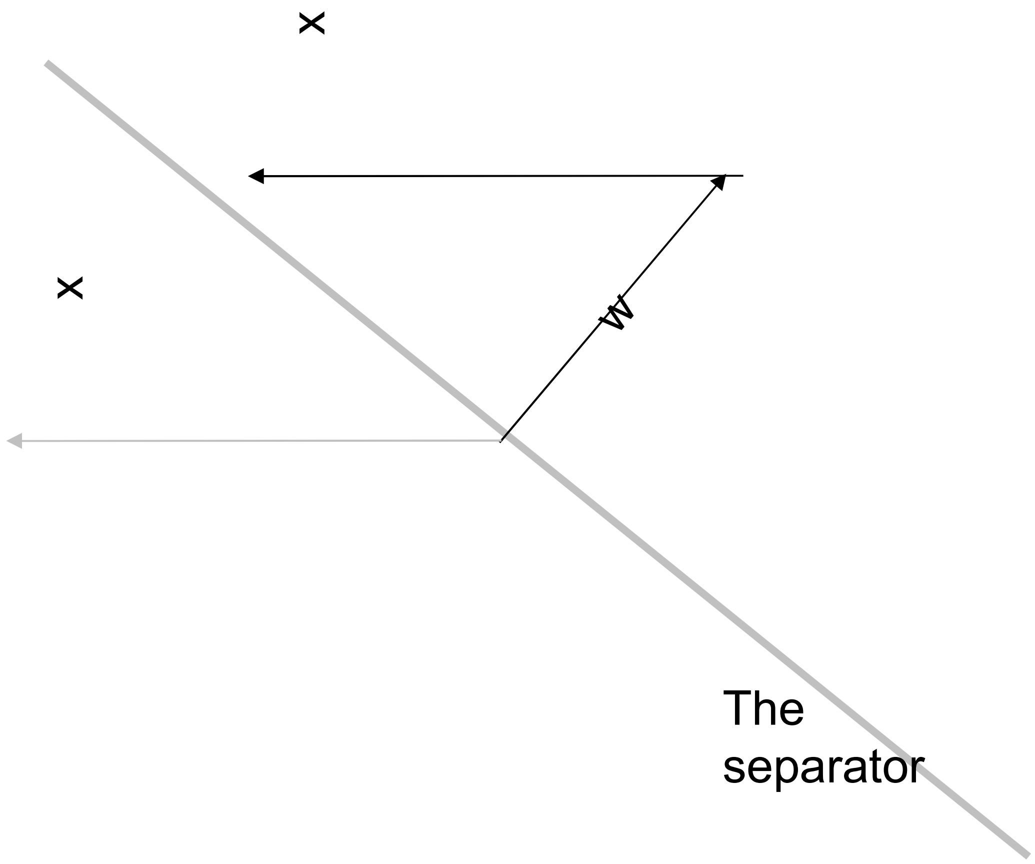
- How does it learn?
  - Each datapoint is annotated with class value 1 or 0
  - Function returns 1 or 0 depending on which side of the separator the point lies
  - Calculate difference between actual and desired output
  - Multiply input vector by this delta and add it to the weight vector
  - Given sufficient iterations the separator will find a solution

Perceptron update



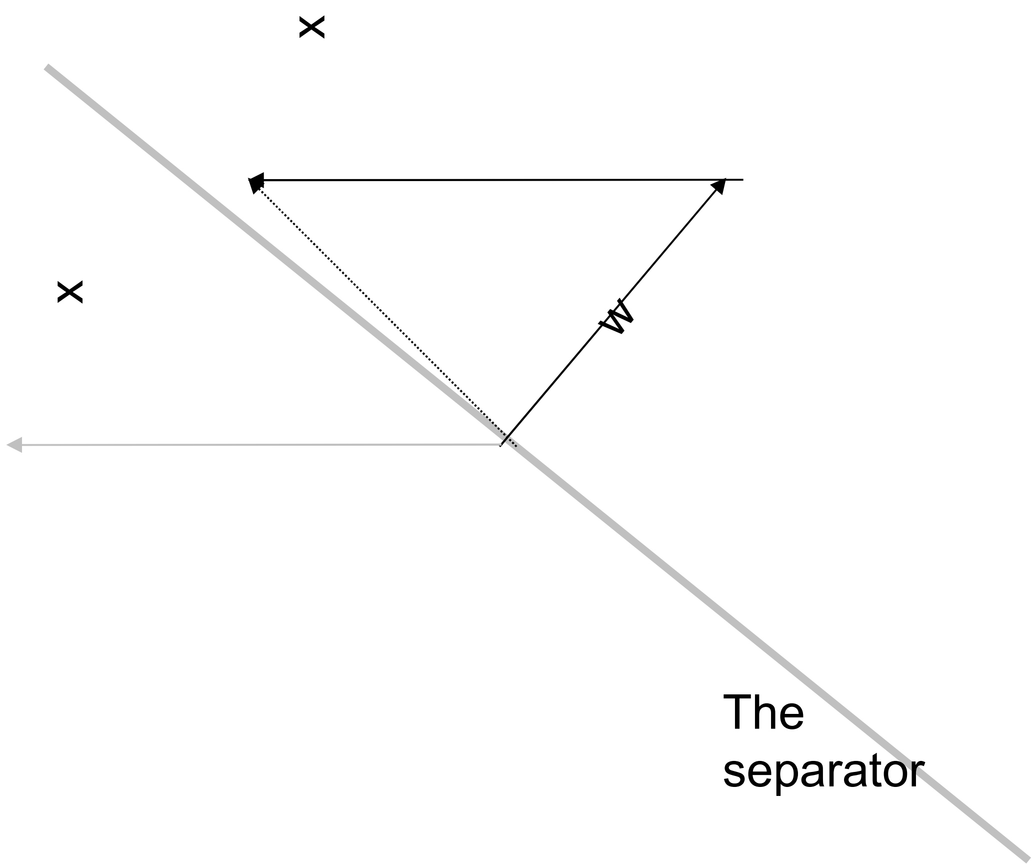
- Dot product is negative, so f=0
- But x is a positive example!
- Oh no! Must update

Perceptron update



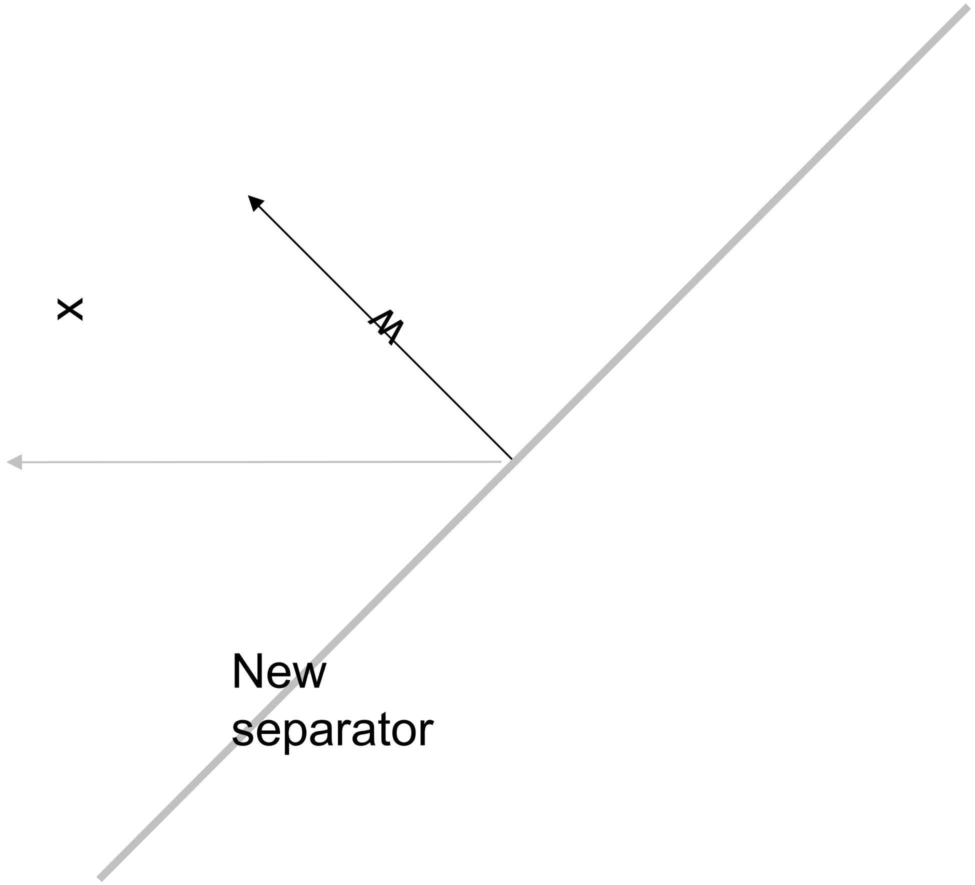
- x class is 1
- f(x) = 0
- w += (1-0)x

Perceptron update



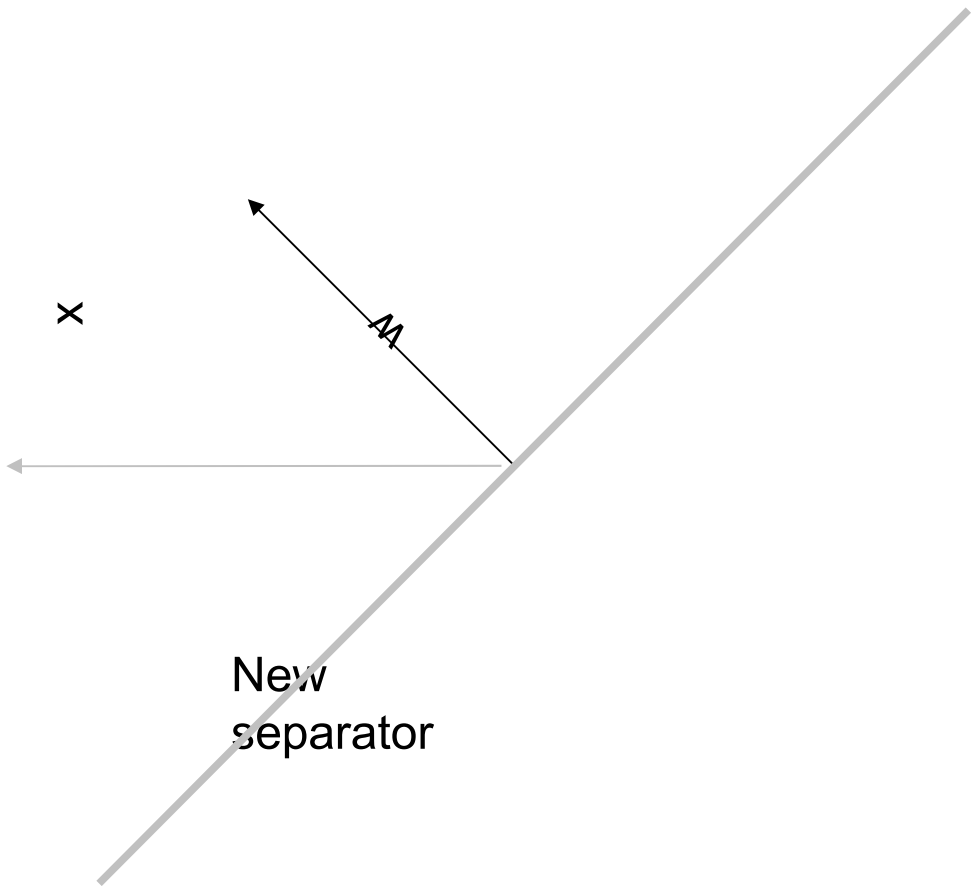
- x class is 1
- f(x) = 0
- w += (1-0)x

Perceptron update



- x class is 1
- f(x) = 0
- w += (1-0)x

Perceptron update

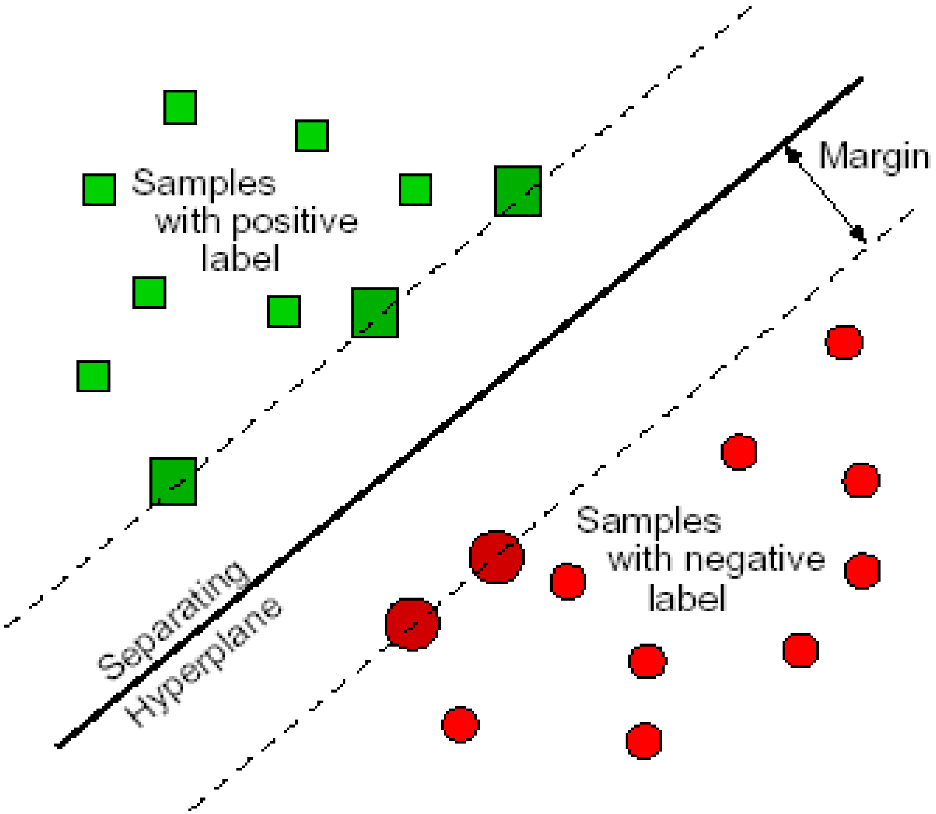


- Now x is on the right side of the separator!

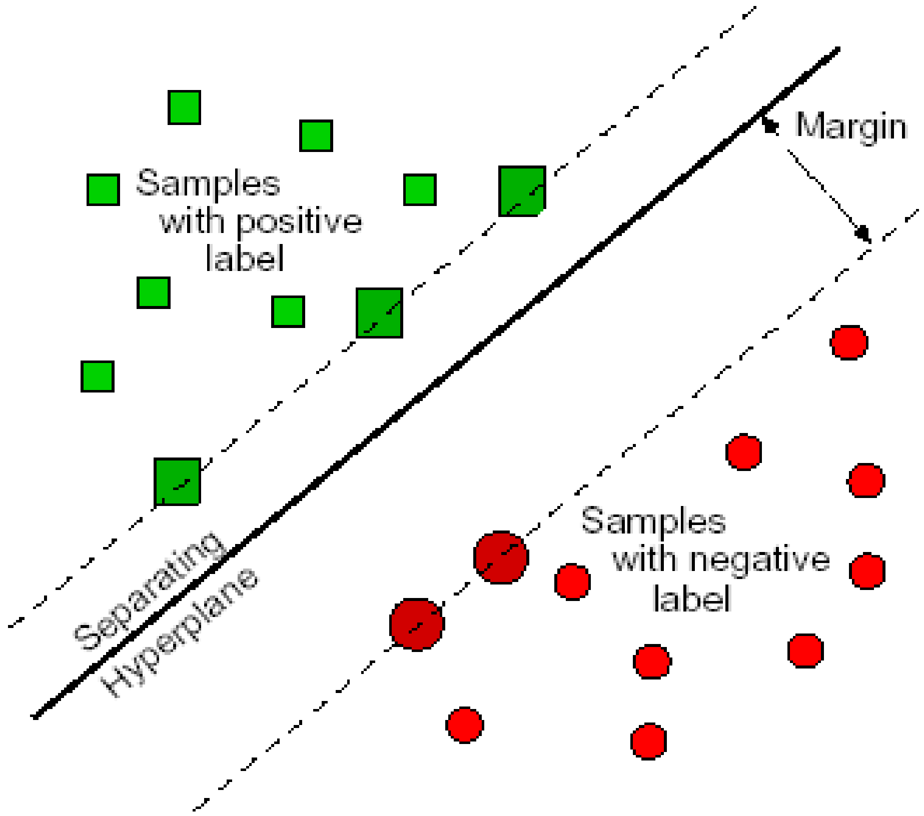
Perceptron with Uneven Margins

- Both Perceptron and SVM implement "uneven margins"
- (PAUM stands for Perceptron Algorithm with Uneven Margins)
- This means that it doesn't position the separator centred between the points, but more towards one side

Even Margins



Uneven Margins



Why Uneven Margins?

- In NLP the datasets are often very imbalanced
- For example if you are finding instances of “Person”, you will have very many words that are not people and only a few that are
- Uneven margins may help with this
- Y. Li, K. Bontcheva, and H. Cunningham. Using Uneven Margins SVM and Perceptron for Information Extraction. Proceedings of Ninth Conference on Computational Natural Language Learning (CoNLL-2005), pp. 72-79. 2005.

Some Other Algorithms

- Batch Learning PR also includes the following from Weka
  - Naïve Bayes
    - Uses Bayes' theorem (probabilities) to determine the most likely class given attributes and training corpus
  - K-Nearest Neighbour
    - Determines class of a point based on k training points positioned geometrically closest to it
  - C4.5 (decision tree)
    - Makes a series of binary decisions that determine the class of a point based on its attribute values (e.g. “is string length > 3?”)